Journal of Educational Research & Policy Studies
Fall 2009, Vol. 9, No. 2, pp. 141 - 163

NORMES

Student Achievement Data and Findings, As Reported In Math and Science Partnerships' Annual and Evaluation Reports

Robert K. Yin
COSMOS Corporation

A primary feature of the Math and Science Partnership Program Evaluation (MSP PE) is the examination of K-12 student achievement changes associated with the National Science Foundation's (NSF) Math and Science Partnership (MSP) Program. This article describes one of three complementary assessments of K-12 student achievement being conducted by the MSP-PE, and consists of a synthesis of student achievement findings reported by the MSP projects themselves (the other two assessments also are described in this volume). The assessment described in this article covers 39 of the 48 MSP project awards made by NSF from 2002 to 2004. Data sources included the MSP projects' annual and evaluation reports submitted to NSF through 2006-07 and research manuscripts developed by the MSPs for presentation at three MSP evaluation conferences. A two dimensional cross-MSP matrix was developed to reveal the disparate research efforts undertaken by the MSPs and present a cross-MSP perspective. The article describes a number of challenges faced by the MSPs as revealed by the current assessment, including: a) many of the MSPs report districtwide data even though the MSPs may not have implemented activities at the district level; b) MSPs that have chosen to define pre-established benchmarks for later comparison to actual performance have not usually discussed any rationale for selecting their particular numeric benchmarks; c) many MSPs report scores for multiple grade levels for both science and mathematics, making an overall interpretation difficult; d) the MSPs should endeavor to identify the amount of professional development that appears to make a discernable difference in student achievement outcomes; and e) most of the evaluation frameworks reported by the MSPs are not poised to go beyond establishing concurrent trends and testing more strongly the actual efficacy of an MSP's activities.

Introduction: Three Complementary Studies on K-12 Student Achievement Trends Associated with the Math and Science Partnership (MSP) Program

The MSP Program consists of a series of separate project awards to individual math and science partnerships (MSPs) made by the National Science Foundation (NSF). Each award went to a different institution of higher education (IHE), which was required to partner with one or more school districts and their numerous schools (see Scherer, this volume). The ensuing partnership was expected to fulfill several objectives. The most prominent of these objectives, as stated in the MSP Program's solicitations issued by NSF, is as follows: "MSP projects are expected to raise the achievement levels of all [K-12] students and significantly reduce achievement gaps in the mathematics and science performance of diverse student populations" (e.g., NSF 09-507, p. 2). Assessing the relationship between the MSP Program and K-12 student achievement therefore has served as a critical part of the Math and Science Partnership Program Evaluation (MSP-PE).

The importance of this evaluation function has led the MSP-PE to design and conduct three complementary assessments. The present article is one of the three assessments focused on student achievement. The three are complementary in that each follows a different research design and uses a different source of data. Each has different strengths and weaknesses, but all are aimed at assessing the potential association between the MSP Program and K-12 student achievement.

The first of the three assessments (Dimitrov, this volume) analyzes student achievement trends based on extensive school-level data submitted into the MSP Program's Management Information System (MSP-MIS) by the MSP projects' schools and districts. The analysis is limited to trends at the MSP sites only, thus far covering three academic years (2003-04 to 2005-06). Among other topics, the MSP-MIS database permits inquiries into the direction of the multi-year trends as well as the association, if any, between the amount of a school's teachers participating in MSP-supported activities and the school's later student achievement. However, because the data source has no information about non-MSP schools or sites, no comparative framework is possible.

The second of the three assessments (Wong, Boben, Kim, & Socha, this volume) compares student achievement patterns between MSP and non-MSP schools, based on school-level data from state education agencies' Web sites. The study's distinctive strength is its careful demographic matching of MSP and non-MSP schools. However, the time-consuming nature of accessing data from individual state Web sites has limited the inquiry to only a small proportion of the MSP projects thus far. Moreover, the study has not yet been able to ascertain the extent of "MSP-like" activities taking place at the non-MSP schools. The non-MSP schools, though accurately matched demographically, therefore cannot be assumed to be "no-treatment" sites.

The present and third assessment, represented by the present article, consists of a synthesis of student achievement findings reported by the MSP projects themselves in

their annual and evaluation reports submitted to NSF through 2006-07. Each MSP has an individual evaluator working with the partnership. These evaluators have employed their own (different) evaluation designs to investigate any relationship between their MSP's activities and student achievement outcomes. In reviewing the MSPs' and evaluators' reports, this third assessment assumes the nature of a research synthesis or secondary analysis, across the MSP projects.

The three assessments all focus on the MSP Program as a whole. They do not attempt to evaluate the individual MSP projects. Overall, the tri-study effort is appropriate for evaluating a program as broad and diverse as the MSP Program. The program does not suggest, much less require, that projects implement any pre-specified educational practices, professional development models, or other uniform initiatives. Instead, each project has been free to devise its own agenda, to meet its own local needs, for improving K-12 student achievement in mathematics and science.

For instance, some MSP projects have undertaken comprehensive activities covering both mathematics and science, across all grade levels. Other projects have limited themselves to either mathematics or science and to specific elementary, middle, or high school grade spans. As another example, some MSP projects have provided large amounts of inservice training to existing K-12 teachers of mathematics and science, whereas other projects have provided less (and different) inservice training but have reorganized the preservice programs at local universities for the purpose of training new teachers.

The varied efforts across the MSPs, combined with their different if not unique partnership configurations, make it impossible to use a single evaluation study or study design to evaluate the relationship between MSP activities and student achievement outcomes (Yin, 2008). Trying to implement a single experimental design or a single evaluation study would lead to either an overly narrow or a superficial depiction of the MSP Program. The alternative has been to pursue three separate assessments, a later goal being to conduct a research synthesis to determine whether the three assessments produce converging findings about the MSP Program. Depending on the nature of the final data, the synthesis can follow more traditional methods (e.g., Cooper, 1998) or employ meta-analytic techniques (e.g., Cooper & Hedges, 1994).

Synthesis Procedure

Scope of Inquiry

The present assessment covers 39 of the 48 MSP project awards made by NSF from 2002 to 2004.[1] The awards were made in annual cohorts, so that the MSP projects were reporting about their third, fourth, or fifth year of work in the reports reviewed for this synthesis. Much of the evaluative data in these reports are based on the work of the evaluators affiliated with each MSP. Nevertheless, the reports are official submissions by the MSPs and potentially suffer from the known limitations of self-reported data.

The synthesis draws from an analysis of the latest available project reports,[2] including both annual and evaluators' reports, as well as from research manuscripts related to presentations at the MSP Program's three evaluation conferences.[3] The data and analyses provided by the individual MSPs were then compiled into a cross-MSP matrix, discussed next. The Appendix to this article contains brief summaries about the nature, status, and findings about student achievement reported by the individual MSP projects, based on the various sources.

Two-Dimensional, Cross-MSP Matrix

To represent the disparate research efforts by each MSP but nevertheless to create the needed cross-MSP perspective, the synthesis characterized every MSP's reported status according to two dimensions, detailed in Table 1:

a) *Evaluation framework*: whether and how an MSP was establishing any comparative framework for interpreting the student achievement outcomes; and

b) *Direction, if any, of findings regarding student achievement trends*: Whether the MSP had started analyzing the data, and if so, whether the data represented mixed, positive, or negative trends over the course of the MSP's award period to date.[4]

The first dimension shown in Table 1, *evaluation framework*, had five categories:

1) "none" (no data had been collected or no framework yet established);
2) "MSP sites only" (the framework had no comparative perspective);
3) "MSP compared to pre-established benchmark or to district- or state-wide averages" (a pre-established benchmark might be an MSP's stated goal that scores would increase by five percent each year; a comparison to either district or statewide averages would reflect an MSP's goal of exceeding these averages);
4) "distinctive within-group comparisons" (see text below for further description of these designs); and
5) "MSP and non-MSP groups compared" (the framework included data from comparison groups of non-MSP classrooms, schools, or districts).

It should be noted that the first and second categories yield no information about any possible association (much less attribution) between an MSP project's work and student achievement. Similarly, the third, fourth, and fifth categories only begin to test such an association, with the third category still being a fairly weak framework.

Table 1
Student Achievement Trends Reported by Cohort I, II, and III MSPs (n = 39)*

Evaluation Framework for MSPs' Analyses	MSPs' Interpretation of *Direction of Findings* (1) No Analysis Yet	(2) No Notable Differences or Mixed Pattern	(3) More Positive than Negative Findings	(4) More Negative than Positive Findings	Total
(1) None	10	0	0	0	10
(2) MSP Sites Only	2	2	3	0	7
(3) MSP Compared to Pre-established Benchmark or to District- or Statewide Averages	0	5	5	0	10
(4) Distinctive Within-group Comparisons	0	3	1	0	4
(5) MSP and non-MSP Groups Compared	2	4	2	0	8
TOTAL	14	14	11	0	39

Note. *The analysis covers the awards to 48 MSPs made by NSF from 2002 to 2004, covering "comprehensive," "targeted," and "institute" types of MSPs. Of the original 48 awards, two were discontinued and seven were "institute" awards that were not included in the present analysis. During the same three-year period, the program also supported 28 other awardees that are not MSPs but that are conducting research, evaluation, and technical assistance activities. These 28 awardees also fall outside of the present analysis. Finally, starting in 2006, the program has since made additional MSP awards that were too new to be included in the analysis. Source: MSPs' Annual and Evaluators' Reports.

The second dimension in Table 1, *direction of findings*, had four categories:

1) "no analysis yet" (whether data had been collected or not, the MSP had made no tallies or observations about the data);

2) "no notable differences" (the MSPs interpreted their own findings as reflecting either no differences or mixed results—i.e., improvement for some academic subjects but not others; or for some grade levels, but not others);

3) "more positive than negative scores" (all or most of the scores favored a positive assessment of the MSP's efforts to date); and

4) "more negative than positive scores" (all or most of the scores favored a negative assessment of the MSP's efforts to date).

In regard to this second dimension, it should be emphasized that the "direction of findings" data represents an MSP project's own interpretations, as stated in its annual or evaluation reports, not on any independent re-analysis or re-interpretation of the MSP's original data.

Remainder of This Article

The remainder of this article has two sections. The first presents the pattern of findings reported by the MSPs. The second comments about the findings and the MSPs' methods for assessing student achievement performance and trends.

Findings Reported by MSPs

Table 1 shows the distribution for the 39 MSP projects under review, according to the two dimensions and their various categories, producing a matrix of 20 cells: The five rows in the matrix represent the *evaluation framework* dimension, and the four columns represent the *direction of findings* dimension. [When relevant, the discussion below cites the designated rows (1, 2, 3, 4, or 5) or columns (1, 2, 3, and 4) in Table 1.]

Evaluation Framework

Examining the *evaluation framework* dimension, the reported frameworks are consistent with an earlier report (Yin, 2007) and are still not especially strong. The five rows are arrayed in ascending order, from weaker (Row 1) to stronger (Row 5) frameworks.

The distribution in Table 1 shows that:

1) 17 (or 44 percent) of the 39 MSPs reported no framework or that they were analyzing "MSP sites only" (Rows 1 and 2), while

2) 10 (26 percent) of the 42 reported comparing their MSP scores with some external benchmark (Row 3), and

3) the remaining 14 (31 percent) reported using a more formal research design involving either some potentially "distinctive within-group comparison" or a non-MSP comparison group (Rows 4 and 5).

Although the total number of MSPs had risen from 34 covered in the earlier update to the 39 in the present one, and although the present analysis took place one year after the earlier update, the proportion of MSPs in these rows still closely mimicked the proportion in the earlier update, which had reported 47, 21, and 32 percent respectively for the same three categories (Yin, 2007).

Among the five types of frameworks, one type–"distinctive within-group comparisons" (Row 4)—has been listed separately because it can include pertinent comparisons by taking into account different amounts of "exposure" to MSP activities, on the part of an array of sites in the same MSP. The row includes what might be

considered frameworks that are stronger than those in Rows 1, 2, and 3, even though they are based on *within*-group comparisons. This is because an MSP-only group can nevertheless be subdivided into two (or more) subgroups. One subgroup can then be exposed to one part of an MSP's activities (e.g., one strand of a curriculum), and another subgroup exposed to a different part (e.g., a different strand). The within-group design then examines whether, if the MSP activities have any potency, the first subgroup performed better when tested on the first part (strand) but not the second, and whether the second subgroup performed better when tested on the second part (strand) but not the first (see Table 2).[5]

Table 2
A Helpful Within-Group Design

School	% of Students Achieving Proficiency in Grade 5 Science, December 2005			
	Strand 1	Strand 2	Strand 3	Strand 4
CEHE	59.7	49.0	42.5	56.4
CENT	48.3	51.8	36.0	53.3
COSP	54.6	49.1	54.9	54.7
EAST	55.5	46.7	42.2	52.3
EBEN	52.0	48.1	44.5	53.1
HARM	64.4	54.0	43.8	62.9
LNES	60.6	**60.5**	51.6	**67.5**
LAES	61.5	54.1	57.9	62.3
MONT	53.3	46.7	41.7	46.0
TMO	72.4	60.8	48.9	65.9
SCOT	58.6	54.5	39.5	46.0
SHAR	**71.1**	51.0	**56.9**	54.7
SHEP	62.5	58.4	54.8	57.7
TCES	51.7	46.6	40.1	50.1
TRES	56.8	53.4	44.8	56.7
UGES	51.1	57.9	56.7	57.1
WHES	57.6	58.4	53.8	65.2
DISTRICT	58.2	52.9	47.5	57.2

Source: MSP's Annual and Evaluator's Reports.

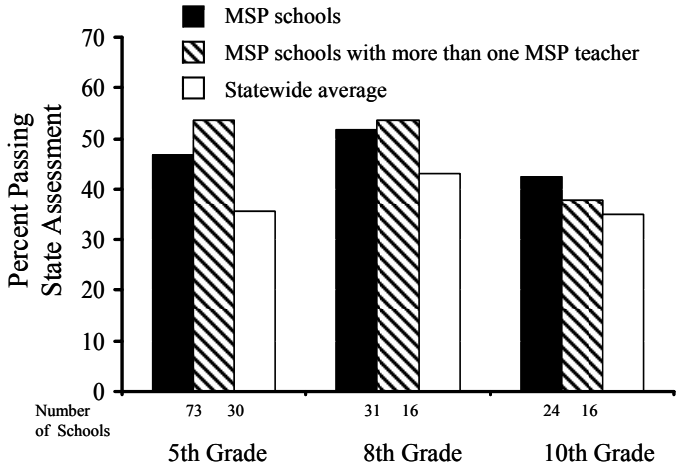
Another within-group strategy among the MSPs has been to compare student achievement trends among classrooms or schools receiving varying degrees of MSP exposure. Because such exposure varied among MSP participants, an analysis could explore whether greater exposure was associated with better student achievement, compared to participation involving less exposure. All participants, however, were MSP participants, and in this sense the framework remained a within-group framework.

For example, Figure 1 presents the data from one MSP that compared: a) schools with one MSP-trained teacher with b) schools having more than one MSP-trained teacher, and with c) statewide averages. A fuller rendition of this design, attempted

by a couple of other MSPs and also exemplifying the within-group efforts in Row 4, called for examining the potential correlation between different amounts of MSP exposure with differing degrees of student achievement. The few MSPs that were able to implement this design found no correlation and hence were categorized as having found "no notable differences."

In contrast, the MSP-non-MSP designs in Row 5 made explicit attempts to collect data from sites totally uninvolved with the MSP's activities. However, except for one MSP, none of the other MSPs defined their non-MSP groups in an especially compelling manner. Whereas the MSP sites were those whose teachers or students had participated in an MSP's activities, the non-MSP sites were simply the neighboring classrooms, schools, or districts that were not participating. Only rarely did the MSPs discuss the possibilities of self-selection between the two groups, and only rarely did the analyses control for other differences between the two groups.

As an exceptional example, only one MSP completed its analysis of results when classrooms were randomly assigned to "treatment" and "non-treatment" conditions. Figure 2 shows that the MSP found no statistically significant differences in the student achievement scores between the two groups; however, the figure also shows that the MSP found no differences in the instruction provided to the two groups. The lack of instructional differences came somewhat as a surprise, because the teachers of the treatment group had participated in the MSP's activities while those of the non-treatment group had not.[6]



Source: MSP's Annual and Evaluator's Reports

Figure 1. MSP schools' scores higher than statewide averages in science, spring 2006.

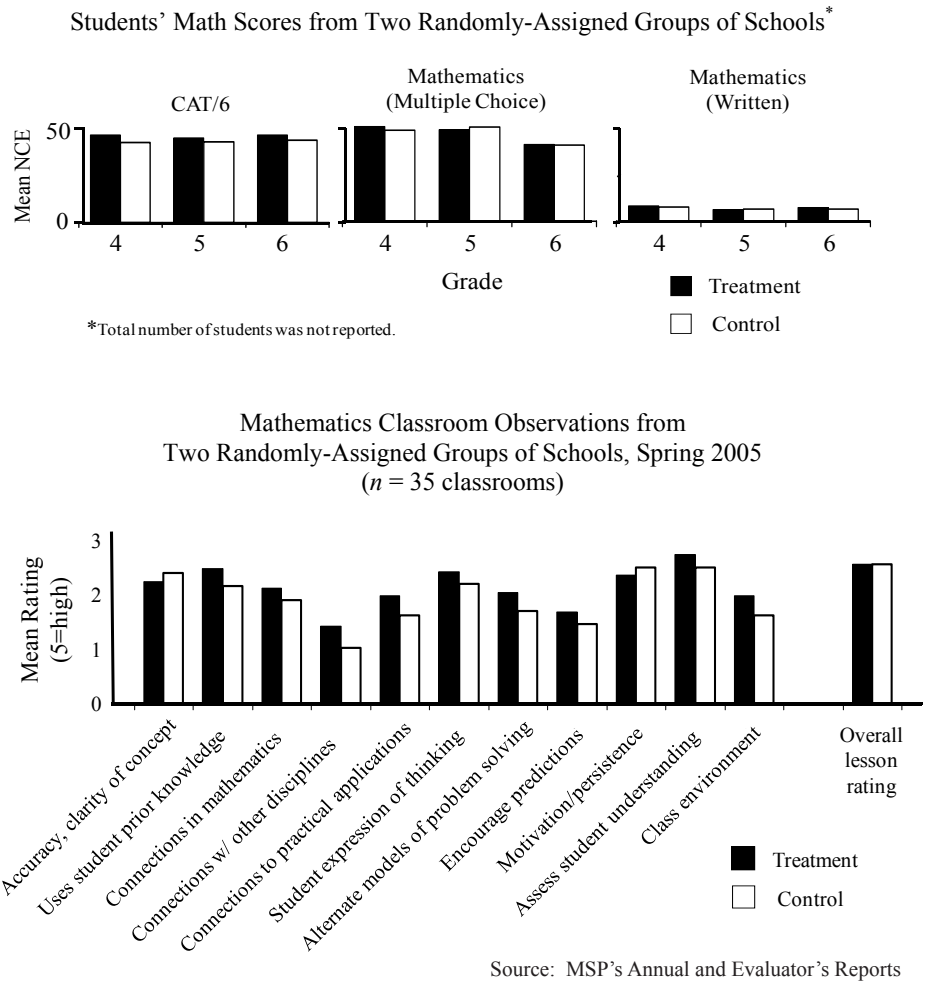


Figure 2. An MSP that randomly assigned classrooms to "treatment" and "non-treatment" conditions.

Other MSPs in Row 5, as previously mentioned, merely identified the non-MSP comparison group as a neighboring classroom, school, or district that had not participated in the MSP's activities. One such MSP was slowly scaling up its work within a participating district and found that its non-MSP group (the remainder of the schools that had not been scaled up) was slowly diminishing.

Direction of Findings

The first important observation about the *direction of findings* is that 14 (31 percent) of the 39 MSPs reported "no analysis yet" (see Table 1, Column 1), while

the remaining 25 (69 percent) of the 39 MSPs reported some analytic findings (see columns 2, 3, and 4). Again, although the total number of MSPs has increased, this proportion is strikingly similar to that in the earlier update (Yin, 2007), which reported 32 and 68 percent respectively for the same two categories.

Somewhat disappointing were the 14 of the 39 MSPs that were still reporting, at this rather late stage of their work, no analysis of student achievement trends. These 14 MSPs only included three projects that were from the later Cohort III awards, so the absence of analysis mainly occurred among MSPs that were already in their fourth or fifth year of work. Several of the MSPs had collected baseline data, but they had not yet collected or analyzed data covering a later period of time. In a few cases, the MSPs were befuddled by a change in their state's assessment test, which in their view made the earlier assessment data impossible to use. These MSPs then reported that they were waiting to obtain at least two years' of scores on the new test, before attempting any analysis. However, one of these MSPs admitted that its coming analysis therefore could not include the desired baseline year of its work.

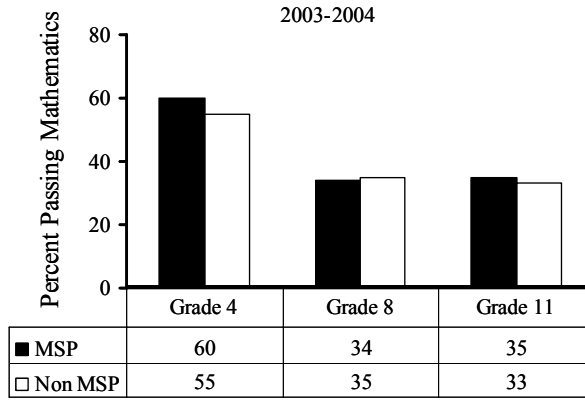
Among the 14 MSPs that had not yet done any analysis, Table 1 shows that ten also had not reported any analytic framework. Assuming the accuracy of the MSPs' reports, the lack of a framework or evaluation design, still a relevant condition during the MSPs' fourth or fifth year of award, may pose an additional challenge when these MSPs eventually pursue any analysis of student achievement data.[7]

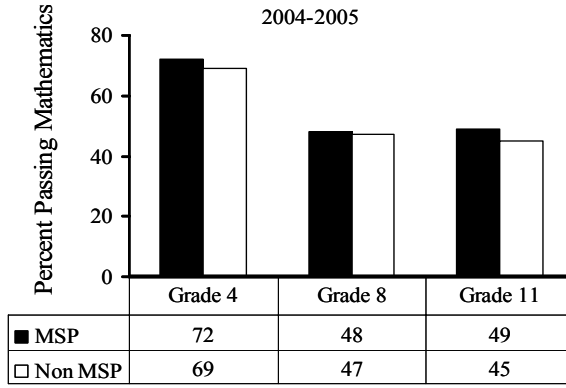
Among the 25 MSPs reporting some analytic findings in Table 1, 14 reported no notable differences or mixed patterns, 11 reported more positive than negative findings, and none reported more negative than positive findings. (The overwhelmingly positive slant poses renewed caution in using self-reported data.) Of the 14 MSPs that had found either no notable differences or a mixed pattern of results, Figure 3 contains the data reported by one MSP showing the similarity of scores between MSP and non-MSP schools and that therefore resulted in the MSP's reporting no notable differences. Similarly mixed patterns were reported by MSPs who might have tested two or more grade levels or two or more academic subjects, or both, but who then found some scores improving and others not.

Figure 4 presents the data illustratively reported by one of the MSPs finding more positive than negative results. The MSP has focused on providing professional development to teachers of mathematics in partnering middle and high schools, and in particular on the performance of the Hispanic students who comprise 67 percent of the enrollment at these schools. The MSP claims that the Hispanic students have attained greater enrollment in higher-level mathematics courses, compared to statewide and county averages. More important, and as shown in Figure 4, the students performed better on the 10th grade state assessment in mathematics in the three years (2003-04 to 2005-06) following the start of the MSP's activities (2002-03).

Frameworks and Direction of Findings Combined

Observing the overall pattern of frequencies in Table 1, and if one accepts that the Rows are arrayed in a sequence from less to greater analytic strength, the distribution of data for the 39 MSPs suggests that the MSPs using the weaker frameworks (Rows





Source: MSP's Annual and Evaluator's Reports.

Figure 3. Comparison between students who attain at proficient and advanced achievement levels (passing performance) on the mathematics assessment tests in MSP and non-MSP schools for 2003-04 and 2004-05.

2 and 3) had a slight tendency to report more positive findings in Column 3, compared to "no notable differences" in Column 2, in contrast to the MSPs using the stronger frameworks (Rows 4 and 5).

Although the tendency was only slight, a possible explanation for the preceding pattern starts with the observation that achievement scores in every state tend to rise over time (sometimes because of the scoring systems, and not necessarily because of learning gains). Observing trends at the "MSP sites only" would then reflect this rise, but putting the MSPs in a comparative mode (e.g., with non-MSP sites or with a within-group comparison) might diminish the appearance of a distinctive rise related to the MSP sites alone.

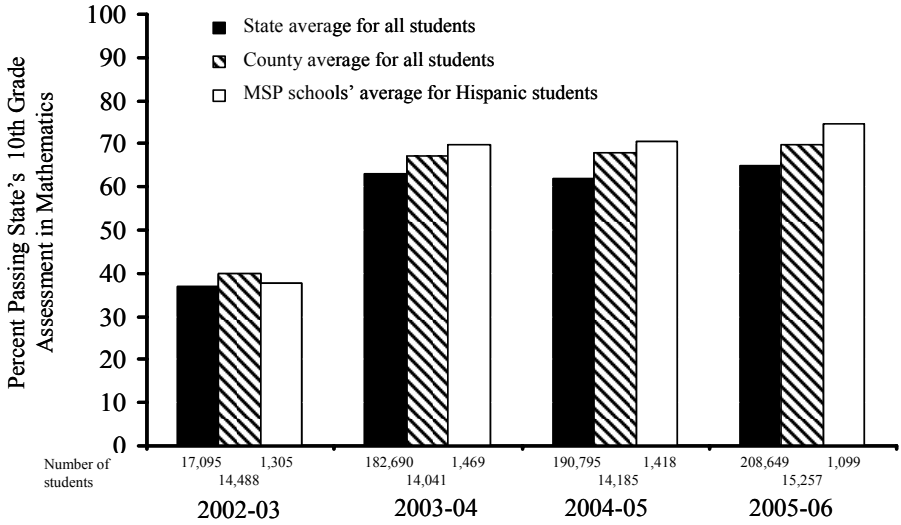


Figure 4. MSP's 10th grade Hispanic students' improvement over state and county averages in mathematics, 2002-03 to 2005-06.

Pattern Across 39 MSPs: Tentative Conclusions

The overall distribution in Table 1 reflects underdeveloped research frameworks and analyses being reported by the MSP projects, regarding their attempts to examine the relationship between their activities and K-12 student achievement. At this rather late juncture, with many MSPs in their fourth or fifth years, 10 of 39 (26 percent) still report no framework and no analysis (Row 1), and 17 of 39 (44 percent) are using frameworks (Rows 2 and 3) that have neither a distinctive within-group nor a non-MSP comparison.

In addition, but not revealed by the distribution in Table 1, only a few of the MSPs have reported using any statistical methods to analyze their data. Such methods are commonly found in evaluations of complex educational programs (e.g., Datnow, Borman, Stringfield, Rachuba, & Castellano, 2003; Supovitz & Taylor, 2005; Yin, Schmidt, & Besag, 2006). Yet, the MSPs have mainly reported their data descriptively. When they report positive or negative differences, such reports are typically based on an author's observations, and not on statistical tests that might determine the strength of any differences. One exception to this cross-MSP tendency is illustrated by an MSP project whose preliminary findings are presented in Table 3. That project is continuing to acquire updated data and to refine its regression models, and its analysis potentially represents the kind of work to be emulated by other MSPs.

Along the same lines, only a few of the MSPs, even those using the stronger frameworks, have reported any attempts to control for demographic and other artifactual conditions when comparing MSP sites to performance by other groups. The

Table 3
Findings From an Analysis Using Fixed-Effect Regression Models (All Schools in State, n = 1,041)

Student Achievement Outcome (No Grade Levels Given)	Time Period	Related to MSP Participation by School*
Change in Math Proficiency Scores	2001-04	Yes, statistically significant
Change in Science Proficiency Scores	2001-04	No
Change in Reading Proficiency Scores	2001-04	Yes, statistically significant

Note. *School characteristics used as covariates: racial/ethnic composition; percentage of students eligible for Free and Reduced-Price Lunch; teacher-student ratio; source: MSP's Annual and Evaluator's Reports.

project depicted in Table 3 again serves as one exception (see the footnote in the table that identifies the covariates used in the regression models).

Conclusions about the MSPs' Existing Assessments of Student Achievement

This synthesis covers ongoing assessments by awardees in NSF's MSP Program. The midstream status of the assessments precludes any conclusions about student achievement. However, the synthesis does lead to several conclusions about the challenges faced by the MSPs.

First, many of the MSPs have reported districtwide data, although the MSPs may not have implemented their activities on a districtwide basis. A similar situation can exist at the school level, where the MSPs' reports may have reported aggregate school-level data, even though a given MSP's activities only may have involved some but not all of the classrooms in the school. In either situation, scale-up may still be occurring, but until fully scaled, the MSPs may need to match more closely their scope of achievement data with the venues in which the MSP activities have taken place.

Second, those MSPs that have chosen to define pre-established benchmarks for later comparison to actual performance have not usually discussed any rationale for selecting their particular numeric benchmark. For instance, the MSPs do not discuss whether such benchmarks as "improving performance by five percent each year" might be too conservative or overly ambitious. The latter could exist if a school already had been improving by substantial percentage points for the preceding years. Where benchmarks are to be used, some discussion and rationale for the selected cut-points would be helpful.

Third, many MSPs report scores for multiple grade levels and for both science and mathematics assessments. MSPs in this situation might want to consider setting another type of benchmark: whether all scores are expected to improve or whether only one or a few are.

Fourth, a promising within-group design examines the relationship between different

degrees of MSP participation (e.g., different amounts of professional development hours) with student achievement outcomes. The hypothesis is that greater participation should be correlated with stronger outcomes. However, such analyses need to be preceded by some pilot study demonstrating the threshold number of hours needed to produce a measurable outcome in the first place. Absent the identification of such a threshold, the possibility remains that the observed professional development hours, in the later correlative analysis, all may fall either under or over the threshold, thereby explaining the lack of any correlation. More generally, the MSP projects should try to establish, through small-scale or pilot testing, the amount of professional development that appears to make a discernible difference in student achievement scores, regardless of whether the later assessment frameworks are based on a correlational analysis or any other analysis.

Finally, most of the evaluation frameworks reported by the MSPs are not poised to go beyond establishing concurrent (or associational) trends and testing more strongly the actual efficacy of an MSP's activities–that is, whether the activities actually had influenced the student outcomes. Because of the large size of the MSPs and the number of students and teachers involved, possibly the MSPs could try to implement some small-scale research, focusing on a few classrooms or schools, that would nevertheless use more robust research designs to assess efficacy.

The sum of these conclusions suggests room for improving the evaluation efforts among the individual MSPs in addressing student achievement outcomes. In particular, the ongoing as well as future efforts should give more attention to the technical design of their inquiries, which can include both qualitative and quantitative methods when trying to relate actions from complex educational programs with student achievement outcomes (e.g., Yin & Davis, 2007).

At the same time, the evaluators have had to struggle with making their assessments at an earlier stage than is usually found in education research. For instance, most published studies of large-scale student achievement trends contain data that are usually at least five years old. Moreover, the MSP evaluators' ability to obtain the needed data from state education agencies –and especially student-level data – may have become more difficult with states' pre-occupation with their own reporting requirements under *No Child Left Behind*. Thus, if programs like the MSP Program expect to deal with student achievement findings in the timely manner still being pursued by the evaluators, the program may want to consider encouraging stronger collaboration among the MSPs, their evaluators, and the state assessment agencies.

Appendix

K-12 Student Achievement Reported by Cohort I, II, and III MSPs ($n = 39$[8]):
(Brief Descriptions)

Cohorts I And II

1. Awaiting 2006-07 data, to compare with 2005-06 baseline. Actual baseline would have been 2004-05 or earlier (new instructional practices started in 2004-05).

However, state changed its mathematics assessment test in 2005-06.

2. Award ended by mutual agreement between awardee and NSF.

3. Data show improvements for five districts from 2002-03 to 2003-04, for both math and science (grade levels not given), with performance in second year also exceeding state's pre-established benchmark of 70 percent proficient. Only defined one of the five districts as having a non-MSP comparison district. Data showed that this single comparison district also improved in both math and science, and it also exceeded the state's benchmark in science. Separate research study uses 2000-01 and 2003-04 data and shows significant relationship between schools' MSP participation and math, but not science scores, compared to all other schools in the state; interpretation clouded by also finding significant relationship with reading scores, though MSP had no reading-related activities (Craig, 2006).

4. District's "percent passing" scores on grades 3-11 mathematics and grades 5, 10, and 11 science all improved for each of two 2-year intervals (2002-03 to 2003-04 and 2003-04 to 2004-05), even though the state raised its standards for "percent passing" each year. The positive trends continue through 2005-06, and White-Hispanic gaps also have been decreasing strongly. However, no comparisons are made, either to statewide averages or to any comparison group, and no attempt is made to link the district's scores to specific MSP activities or the time of those activities. Report for 2005-06 contains no student achievement data.

5. Finds elementary school and middle school students improve significantly in math, compared to ELA (but ELA gains greater in grades 2 and 3), for classes exposed to the MSP's three-prong activity consisting of curriculum (pacing) guides, quarterly assessments aligned with state assessments, and PD to help focus on needed instruction (Hyde, Mann, Manrique, & Shanahan, 2005; Shanahan, Mann, & Manrique, 2006).

6. Original MSP reorganized and now has an official start date of 1/06. Subsequent reports and an MSP-PE site visit have revealed little progress in analyzing any student achievement data, even though the MSP's first goal is to increase student achievement in the partnering school district.

7. Fourth-Year evaluator's report says that mean achievement scores for three of four SCALE districts "...have varied little from prior to 2003 (before SCALE) and after 2003 (with SCALE)" (Porter, 2006, p. 4). However, the MSP's fourth-year report also indicates highly incomplete scale-up, with the "upper limits" of teacher participation estimated at 67 and 66 percent in two of the districts, 41 in the third, and 10 in the fourth, and with no discussion of effects of teacher turnover or in classroom assignments after the MSP activity took place.

8. Earlier analyses had indicated that MSP schools and non-MSP schools did not differ in baseline (Bravo & Arce, 2006). Separate evaluator's report goes on to show

that the percentage of students at or above proficiency increased between 2003-04 and 2004-05, for math in grades 4, 8, and 11, beyond a 5 percent benchmark set by the MSP. However, there were no significant differences between the gains for the MSP schools and those for the non-MSP schools, "...so the gains cannot be attributed to the MSP's reforms" (Vesperman, Mayer, & Webb, 2005, p. 84).

9. Administered TIMMS to about 200,000 students, grades 3-12, as baseline data. Implementing a complicated randomized field trial with districts assigned to multiple combinations of conditions. Trial continued through 2006-07, so no analysis yet.

10. Multi-year scores for MSP's district show no distinct baseline trends (Walker, Gosz, & Huinker, 2005). HLM regression analysis shows that MSP participation, defined as a bivariate condition only, explains but a small percentage of the variability in student proficiency in mathematics for grades 4 and 8 (partly because socio-economic status and prior achievement already account for a large percentage). (A comparison group would produce a dependent variable that is the difference between two groups, not just the performance of one group alone.) The MSP also tested a path model that produced no clear results, and the model is being modified for testing with future data (Hanssen, 2006). Major challenge is to relate PD and school activities with specific classrooms, then with classroom practices, and then with student achievement.

11. Compares multi-district performance with pre-established benchmark (that 90 percent of the MSP districts will have 75 percent or more of students passing the state assessment in mathematics; there is no state assessment in science); by the end of the MSP's third year, many districts were progressing well at grades 5 and 8 but not at grade 11.

12. Latest update (2002 through 2006) shows continuing positive trends for MSP's four regions, but these are not compared to any benchmark or to state or other averages.

13. Continues to find only small differences in teachers' instruction, between randomly-assigned participating and non-participating schools, and similarly finds no significant differences in grades 4, 5, and 6 math scores between the two groups (Bocian, Torres, Bryant, & Hammond, 2005; Torres, Bocian & Bryant, 2006).

14. Two-year achievement trends show increases in "percent passing" and also reduced number of students in "percent below basic," for nearly all 17 participating districts, from 2002-03 to 2003-04, in 6th and 9th grades math and science (Shama-Davis et al., 2005). Later trends harder to track because of changes in state assessment; still need to analyze trends using districts' end-of-grade tests.

15. Serendipitous use of a district assessment shows that, within the same test, and for the whole grade in two different schools, students perform better on two of five strands related to the MSP's science activities than on other strands, in predicted alternating fashion between the schools, mimicking differences in MSP implementation

by school. However, the MSP has not succeeded in its original assessment plan–either having the appropriate test items incorporated into the state assessment or developing its own assessment aligned with its own science kits and curriculum.

16. Cites mixed baseline achievement trends for years prior to MSP, in participating schools; also may be making comparisons with non-MSP schools and classrooms. Concerned with recent changes in state assessment.

17. Data on student achievement appear to suggest positive improvements in relation to MSP's mentoring activity. However, a difficult-to-interpret numeric table is not accompanied by any narrative, and the results also were not addressed during the site visit.

18. Report shows no particular differences between MSP's nine districts and statewide or regional averages, from 2002-03 to 2004-05 in mathematics, summing all grades 3-11. However, evaluator points out that district with 100 percent teacher participation and most intense MSP involvement improved the most and exceeded the statewide and regional averages. Later update finds increases for 8 of 9 participating districts, although MSP's activities are not necessarily covering extensive portions of each district (Lamm & Sloan, 2006).

19. The evaluation team has not completed any outcome evaluation. Earlier formative analyses showed that performance on the specific grade 7 science strand related to the MSP's PD (inquiry-based science) improved from 2001-02 to 2003-04, but not on the 7 other strands tested, at *one* MSP school (see MSP's Yr-2 annual report and the MSP-PE site visit notes, 1/9/07). Other MSP schools did not have the same pattern, possibly because of high baseline scores at the other schools. In grade 8 mathematics, another school had the highest scores in the state, and the school unexpectedly maintained its high scores for 2003-04, possibly because the MSP had implemented PD in *Connected Math*.

20. Evaluation report claims that MSP-participating students (those in the classrooms of the teachers who participated in MSP) did better than a comparison group in grades 6, 7, 8, and 10 but lower in grades 9 and 11 (specific academic subjects not identified). The report provides no further details and does not present the actual data or analysis.

21. No indication of any plan to collect student achievement data. Targeted, mentee teachers (n = about 500) come from over 15 states, and therefore numerous districts and schools, all participating in on-line professional development for beginning teachers.

22. Award ended by mutual agreement between awardee and NSF.

23. The major MSP activity has been aiding districts to implement *Everyday Math*. State assessment for 2003, but not later years, shows that schools with three or more

pilot teachers (who participated in the MSP's PD) performed significantly better in grades 3 and 6 mathematics than schools with no pilot teachers. Comparison between MSP schools and matched set of non-MSP schools show no differences between 1999 and 2004, for 3rd grade mathematics. The data also show reductions in Black-White but not Hispanic-White gaps at some grades and for some of the participating districts.

24. Evaluator reports finding no cross-sectional correlation between the amount of PD to biology teachers (average of 40 hours per teacher for one year) and student performance on biology portion of host district's science assessment for 2004-05; also no relationship between the amount of PD and change in students' scores from 2004-05 to 2005-06 (Frechtling et al., 2006).

25. With MSP's activities having a districtwide reach, districtwide student achievement improving but not different from rest of state in mathematics, Reductions in achievement gaps unclear (Apaza, Sayler, & Austin, 2005; Saylor & Apaza, 2006). MSP also reports that extent of improvement in student achievement is positively correlated with extent of teacher implementation of MSP's instructional materials.

26. Shows that middle school and high school achievement scores have improved, but with no benchmarks or comparisons; enrollment in math courses has increased but could have been influenced by a concurrent increase in the state's requirements from 2 to 3 math courses for graduation.

27. Most recent results show that the (diminishing) control group outperformed MSP groups on most high school science and math Regents exams (but controls outperformed MSP group on 3 of 7 tests in one district and on 5 of 6 tests in the other). Informal analysis suggests that the more teacher training, the better the performance; the more a student is exposed to multiple teachers with training, the better the performance. Earlier, for 2003-04, students taking courses by MSP-trained teachers performed better on state assessment in grade 8 math and science than students in other classes.

28. Minority students show greater enrollment but mixed changes in failure rate, in MSP's first cohort of three participating high schools, compared to target rates set at outset by the MSP.

29. Hispanic students' enrollment rates and reductions in achievement gap, relative to statewide averages, are meeting the MSP's pre-established benchmarks.

30. Compared test scores for 2002 and 2005 with statewide averages. Three high schools that had MSP participation (averaged 51-115 hours per teacher, over three years) did better than statewide averages in 10th grade mathematics; six middle schools that had mixed MSP participation (averaged 23 to 62 hours per teacher, over three years) also had mixed results compared to statewide averages on 8th grade mathematics (Lee, Baldasarri, & Leblang, 2006).

31. Fourth annual report (06-07) does not discuss plans or designs for analyzing student achievement data. However, evaluators have been conducting two types of analyses, one involving 8th grade state assessment data (comparing students whose teachers did or did not attend MSP's PD) and the other using end-of-course district assessments. Because the MSP helped to design the latter, the data are available by strand, so that closer comparisons can be made between test performance on specific strands and teachers' PD exposure to specific PD topics (such close comparisons cannot be conducted with the state assessment data, which are not available on an item basis). The findings are to be reported in the fall of 2007.

32. Starting to collect districts' student achievement data in 2006-07.

33. Findings show that passing rates for middle school math improved for all but two project schools in MSP's initial years. Although the rates exceeded the MSP's benchmarks that called for an increasing proportion of its schools to reach state proficiency levels, the overall improvement for 7 of 10 MSP districts from 2001 to 2007 were no greater than those for 5 comparison districts or for statewide averages.

34. Multiple comparisons to statewide averages for two cohorts show mixed results, but design of the analysis is unclear.

35. State has no 8th grade science test, so MSP has used own test to establish baseline for whole district.

36. Percent scoring proficient or above, for grades 5, 8, and 10 science, are higher than statewide averages in spring 2006, though the scores had tended to be lower than the statewide averages two years earlier. However, percentage gains from 2004 to 2006 were only greater than those for statewide averages for grade 5. Comparisons are even better when examining subset of schools with more than one MSP teacher, except for 10th grade.

Cohort III

37. Still awaiting results from recent science tests, but no discussion of plan for collecting or analyzing data.

38. August 2006 Regents scores show that MSP's summer school participants performed better than those enrolled in regular summer schools.

39. Still collecting achievement data from multiple districts. Also trying to track course enrollment, analyzing patterns for 2005-06 and 2006-07.

40. Just starting to collect baseline data, but no clear identification of target grades or comparison groups.

41. Has collected baseline data. For schools meeting a criterion level of MSP participation, plan to make later comparisons with matched non-MSP schools selected from the rest of the state, in middle school mathematics.

Endnotes

[1]NSF made 48 MSP project awards from 2002 to 2004, covering comprehensive," targeted," and "institute" types of MSPs. Of the 48 projects, two were discontinued and seven were "institute" awards that were not included in the present analysis. During the same three-year period, the program also supported 28 other project awards that are not MSPs but that are conducting research, evaluation, and technical assistance (RETA) activities. However, the RETAs do not necessarily involve activities directly related to K-12 classrooms or teachers. Therefore, the 28 RETAs also fall outside of the present analysis. Finally, starting in 2006, the program has since made additional MSP project awards that were too new to be included in the analysis.

[2] For three projects, the information came from site visits conducted by the program evaluation team (MSP-PE), because the reports did not cover the MSPs' student achievement work.

[3] These conferences were held in Minneapolis, MN, in September of 2005 and 2006, and in Washington, DC in January 2008.

[4] The annual and evaluators' reports varied in the recency of the student achievement data in their analyses. Most of the reports included student achievement data for 2004-05 and earlier years, whether the MSP was from Cohort I or II. A few of the reports included data for 2005-06 (and earlier years).

[5] In few instances, as with the cited MSP, did the MSPs perform any statistical tests to determine the significance of any differences. Where tests were performed, the results are noted in the tables and figures. Otherwise, the data in the tables and figures need to be recognized as descriptive data only.

[6] One other MSP has reported using a more complex design whereby different participating districts have been randomly assigned to multiple combinations of conditions. However, the implementation of this design was still ongoing during 2006-07, so the analysis of these results will not be available for some time.

[7] Although these MSPs may not have reported about their own student achievement analysis in their annual or evaluators' reports, all had submitted (school-level) student achievement data into the MSP Program's management information system (MSP-MIS). Such annual submissions are a requirement of the MSP Program.

[8] The analysis covers the awards to 48 MSPs made by NSF from 2002 to 2004, covering "comprehensive," targeted," and "institute" types of MSPs. Of the original

48 awards, two were discontinued and seven were "institute" awards that were not included in the present analysis. During the same three-year period, the Program also supported 28 other awardees that are not MSPs but that are conducting research, evaluation, and technical assistance activities. These 28 awardees also fall outside of the present analysis. Finally, starting in 2006, the Program has since made additional MSP awards that were too new to be included in the analysis.

Acknowledgments

This article is one in a series of studies for the Math and Science Partnership Program Evaluation (MSP-PE) conducted for the National Science Foundation's Math and Science Partnership Program (NSF MSP). The MSP-PE is conducted under Contract No. EHR-0456995. Since 2007, Bernice Anderson, Ed.D., Senior Advisor for Evaluation, Directorate for Education and Human Resources, has served as the NSF Program Officer.

The MSP-PE is led by COSMOS Corporation. Robert K. Yin (COSMOS) serves as Principal Investigator (PI) and Jennifer Scherer (COSMOS) serves as one of three Co-Principal Investigators. Additional Co-Principal Investigators are Patricia Moyer-Packenham (Utah State University) and Kenneth Wong (Brown University).

Any opinions, findings, conclusions, and recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Apaza, J., Sayler, B., & Austin, M. (2005, September). *Exploring the relationship between the use of standards-based instructional materials and student achievement in mathematics: A pilot study.* Paper presented at the MSP Evaluation Summit, Minneapolis, MN.

Bocian, K., Torres, R. T., Bryant, M., & Hammond, K. (2005, September). *Evidence-based design from the mathematical ACTS MSP project at the University of California-Riverside.* Paper presented at the MSP Evaluation Summit, Minneapolis, MN.

Bravo, M., & Arce, J. (2006, October). *Teachers' and students' learning with understanding in science and mathematics: Methods and results from PRMSP/ Alacima.* Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature synthesis* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cooper, H. M., & Hedges, L. V. (Eds.) (1994). *Handbook of research synthesis.* New York, NY: Russell Sage Foundation.

Craig, W. (2006). *The effects of professional development programs on educational outcomes in mathematics and sciences.* Unpublished paper, Martin School of Public Policy and Administration, University of Kentucky.

Datnow, A., Borman, G., Stringfield, S., Rachuba, L., & Castellano, M. (2003). Comprehensive school reform in culturally and linguistically diverse contexts: Implementation and outcomes from a four-year study. *Educational Evaluation and Policy Analysis*, *25*, 143-170.

Dimitrov, D. (2009). Intermediate trends in MSP-related changes in student achievement with MIS data. *Journal of Educational Research & Policy Studies,* this volume.

Frechtling, J., Miyaoka, A., Parsad, B., Raue, K., & Zhang X. (2006, July). *Evaluation of the Vertically Integrated Partnerships K-16 Mathematics and Science Partnership*. Unpublished report (contained in the MSP's 2005-06 Annual Report), Westat, Rockville, MD.

Hanssen, C. (2006, October). *The Milwaukee Mathematics Partnership: A path model for evaluating teacher and student effects*. Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Hyde, K., Mann, V., Manrique, C., & Shanahan, T. (2005, September). *Integrating curriculum guides, quarterly benchmark assessments, and professional development to improve student learning in mathematics*. Paper presented at the MSP Evaluation Summit, Minneapolis, MN.

Lamm, H., & Sloan, L. (2006, October). *Focused mathematics professional development results in student achievement gains*. Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Lee, S., Baldassari, C., & Leblang, J. (2006, May). *Year 3 evaluation report*. Unpublished report, Program Evaluation and Research Group, Lesley University, Cambridge, MA.

Porter, A. (2006, October). *Evaluator's report* [System-wide Change for All Learners and Educators (SCALE) at the University of Wisconsin-Madison]. Unpublished report submitted to the National Science Foundation.

Sayler, B., & Apaza, J. (2006, October). *Using data to guide mathematics reform within a K-12 district.* Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Scherer, J. (2009). Understanding the role of partnership configuration in the NSF-MSP Program. *Journal of Educational Research & Policy Studies,* this volume.

Shama-Davis, D., Hart, R., Shelestak, D., Stauffer, C., Peters, P., & Keuchel, K. (2005, June). *Year 3 evaluation report, 2004-05* [Stark County Math and Science Partnership], Bureau of Research Training and Services, Kent State University.

Shanahan, T., Mann, V., & Manrique, C. (2006, October). *Continued integration of curriculum guides, quarterly benchmark assessments, and professional development to improve student learning in mathematics and science.* Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Supovitz, J. A., & Taylor, B. S. (2005). Systemic education evaluation: Evaluating the impact of systemwide reform in education. *American Journal of Evaluation*, *26*, 204-230.

Torres, R. T., Bocian K., & Bryant, M. (2006, October). *Evidence based decision making to support teacher learning in the Mathematical ACTS MSP*. Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Vesperman, B., Mayer, R., & Webb, N. (2005, July). *Puerto Rico baseline results in mathematics and Spanish test scores for school years 2002-04*, Unpublished report, Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison.

Walker, C. M., Gosz, J., & Huinker, D. (2005, September). *Measuring the effect of the Milwaukee Partnership on Student Achievement.* Paper presented at the MSP Evaluation Summit, Minneapolis, MN.

Wong, K. K., Boben, M., Kim, C., & Socha, T. (2009). Comparison of MSP and non-MSP schools in six states. *Journal of Educational Research & Policy Studies,* this volume.

Yin, R. K. (2008). The Math and Science Partnership program evaluation: Overview of the first two years. *Peabody Journal of Education*, *83*, 486-508.

Yin, R. K. (2007, June). *Updated examination of student achievement data and findings reported by MSPs in their 2005-06 Annual and Evaluators' Reports*. Unpublished paper, COSMOS Corporation, Bethesda, MD.

Yin, R. K., & Davis, D. (2007). Adding new dimensions to case study evaluations: The case of evaluating comprehensive reforms. *New Directions for Evaluation*, *113*, 75-93.

Yin, R. K., Schmidt, R. J., & Besag, F. (2006). Aggregating student achievement trends across states with different tests: Using standardized slopes. *Peabody Journal of Education*, *81*, 47-61.